뉴로모픽 하드웨어를 고려한 컴포넌트 모델 최적화 기법

김서연[1], 윤영선[1], 홍지만[2], 김봉재[3], 이건명[4], 정진만[5]
[1]한남대학교 정보통신공학과 [2]숭실대학교 컴퓨터학부 [3]충북대학교 컴퓨터공학과
[4]충북대학교 소프트웨어학과 [5]인하대학교 컴퓨터공학과
sykim.hn@gmail.com ysyun@hnu.kr jiman@ssu.ac.kr
bjkim@chungbuk.ac.kr kmlee@chnu.ac.kr jmjung@inha.ac.kr

A Component Model Optimization Method Considering Neuromorphic Hardware

Seoyeon Kim[1], Young-Sun Yun[1], Jiman Hong[2], Bongjae Kim[3],
Keon Myung Lee[4], Jinman Jung[5]
[1]Dept. of Information & Communication Engineering, Hannam University
[2]School of Computer Science and Engineering, Soongsil University
[3]Dept. of Computer Engineering, Chungbuk National University
[4]Dept. of Computer Science, Chungbuk National University
[5]Dept. Computer Engineering, Inha University

요 약

클라우드 서버를 이용한 IoT 응용 개발은 네트워크로 연결된 하드웨어에 데이터 송수신 지연, 네트워크 트래픽, 실시간 처리 지원을 위한 비용 등의 문제가 발생한다. 엣지 클라우드 기반 플랫폼에서는 이러한 문제를 해결하기 위해 빠른 데이터 전달이 가능하도록 뉴로모픽 하드웨어를 사용할 수 있다. 본 논문에서는 뉴로모픽 하드웨어를 고려하여 스파이킹 신경망 모델을 최적화할 수 있는 알고리즘을 제안한다. 컴포넌트의 구성을 세 가지로 분류하여 인코딩 유형 및 데이터 크기를 자동화하며 모델 성능에 영향받지 않는 파라미터를 추상화하고 실행 환경에 최적화된 모델을 제공하는 것에 초점을 맞추었다. IoT 개발자가 파라미터 입력만으로 컴포넌트를 생성할 수 있고 플로우 기반의 오픈 IDE에 적용하여 AI와 결합된 자율형 IoT 응용을 쉽게 개발할 수 있다.

1. 서 론

최근 IoT 디바이스는 센서 및 액츄레이터, 이동성 유무 및 사양에 따라 이질적인 특성을 보이며 다양한 AI(Artificial Intelligent) 응용을 지원하기 위해 높은 컴퓨팅 능력이 요구되고 있다[1-3]. 또한, 클라우드 서버와 연동하여 서비스 요청 및 처리를 진행하기 위해 대용량 자원 요구가 증가하고 네트워크 처리 응답의 지연이나 프라이버시 등의 문제가 발생하고 있다[4]. 엣지 클라우드 기반 플랫폼은 데이터 실시간 처리, 전송량 감소, 데이터 전송 및 트래픽 문제 등의 해결을 지원하기 위해 라즈베리파이, Google 코랄, Nvidia 젯슨 보드 등을 이용하여 경량 AI 응용을 제공한다[5].

IoT 사물에서 지능적인 컴퓨팅 능력 요구에 따라 신속하게 데이터 처리가 가능하도록 뉴로모픽 하드웨어에 관한 연구도 증가하고 있다. 뉴로모픽 하드웨어는 스파이킹 신경망을 통해 AI 알고리즘과 결합된 IoT 사물의 상황 판단 및 추론이 가능한 자율형 IoT를 지원한다[6]. 하지만 뉴로모픽 하드웨어에서 요구하는 3세대의 스파이킹 신경망은 기존 2세대 인공신경망보다 더 생물학적 뉴런에 가깝도록 모델링 된 것으로 IoT 개발자가 해당 모델을 사용하여 자율형 IoT를 개발하기에는 어려움이 따른다[7].

본 논문에서는 IoT 개발자를 위한 자율형 IoT 응용 개발에 초점을 맞추어 뉴로모픽 하드웨어를 고려한 컴포넌트 모델 최적화 기법을 제안한다. 컴포넌트의 구성을 인코딩 유형 및 데이터 크기를 자동화할 수 있는 적응부, 모델 성능에 영향받지 않는 파라미터를 추상화하는 모델부, 실행 환경에 최적화된 모델을 제공하는 실행부, 세 가지로 분류한다. IoT 개발자는 파라미터 입력만으로 AI 컴포넌트를 생성할 수 있고 플로우 기반의 오픈 IDE에 적용하여 AI와 결합된 자율형 IoT 응용을 쉽게 개발할 수 있다.

2. 관련연구

기존 연구[8]를 통해 뉴로모픽 하드웨어는 제약적인 제원으로 인해 각 하드웨어마다 다른 최적화 모델을 생성해야 함을 확인하였다. 뉴로모픽 아키텍처 기반 FPGA보드 중 terasic의 DE1-SoC와 xilinx의 PYNQ는 표 1과 같이 서로 다른 크기의 자원을 제공하며 뉴런과 차원을 곱한 값은 제한적이다. 두 수의 곱이 각 하드웨어 제원의 상한보다 크다면 각 뉴런 또는 차원의 수가 상한보다 작다고 하더라도 보드에서 실행할 수 없다. 예를 들어 DE1-SoC에서 MNIST 데이터셋을 이용한 모델을 생성

할 경우 인풋 데이터 크기가 28 * 28로 784개의 뉴런이 필요하고, 히든 레이어에서는 최대 뉴런과 차원의 곱인 16K보다 낮아야 하므로 20개의 뉴런을 사용해야 한다. 따라서 인풋 데이터의 크기를 14 * 14로 조정하게 되면 히든 레이어의 뉴런 수는 81개로 증가시킬 수 있기 때문에 더 높은 성능으로 수행할 수 있다. 만약 같은 모델을 PYNQ에서 수행할 경우 28 * 28의 이미지를 사용하면 히든 레이어는 40개를 사용할 수 있으며 이미지의 크기를 14 * 14로 조정하여 사용하면 164개의 히든 레이어를 사용할 수 있으므로 더 높은 성능을 기대할 수 있다.

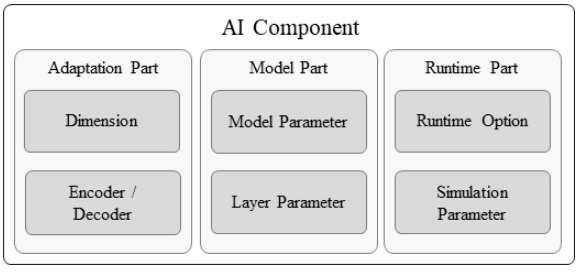
(표 1) 뉴로모픽 하드웨어 FPGA 보드의 자원 크기

Board Name	Neurons(N)	Dimensions(D)	N x D(C_{max})
DE1-SoC	16K	1K	16K
PYNQ	32K	1K	32K

3. 컴포넌트 모델 최적화 기법

3.1 시스템 모델

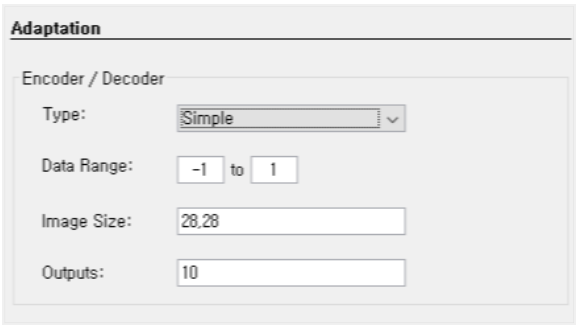
AI 컴포넌트는 그림 1과 같이 3가지 파트를 통해 생성될 수 있다. 3세대 신경망인 스파이킹 신경망은 2세대의 인공신경망과 다르게 스파이크 트레인 형태의 데이터가 전달되어야 한다. 따라서 인코더 및 디코더는 매우 필수적이며 개발자에게 데이터 변환에 대한 이해와 인코딩 정보가 요구된다. 컴포넌트의 적응부는 응용 프로그램과 뉴로모픽 AI 컴포넌트 사이에서 이러한 데이터 가공 및 전달과 같은 데이터 변화에 적응하기 위한 파트이다. 자동으로 스파이크 트레인 형태의 데이터를 전달할 수 있도록 인코더 및 디코더와 데이터 이미지 스케일 등의 파라미터를 자동화하여준다. 컴포넌트의 모델부는 스파이킹 신경망을 구성하는 뉴런 모델에 대한 정의를 위한 파트이다. 뉴로 사이언스에 초점이 맞추어진 스파이킹 신경망 모델에서 생물학적 정보의 지식을 기반으로 많은 파라미터가 필요하지만 정보처리 관점의 AI 및 IoT 결합 응용에서는 성능에 큰 영향이 없는 파라미터를 추상화한다. 컴포넌트의 실행부는 하드웨어 또는 시뮬레이션이 선택되었을 때 수행할 모델이 해당 하드웨어 또는 시뮬레이션에 최적화 시켜준다.



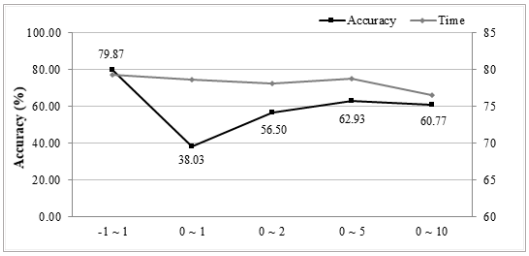
(그림 1) 컴포넌트 생성부

3.2 적응부

컴포넌트의 적응부는 IoT 개발자가 스파이킹 신경망 또는 인공신경망을 이용한 자율형 IoT 응용을 개발하는 것에 초점을 맞추어 스파이크 트레인으로의 인코딩을 자동화한다. 학습 알고리즘이나 모델에 따라 사용할 수 있는 인코딩 방식이 다르기 때문에 사용자에게 추가적인 지식이 요구된다. 또한, 스파이크 트레인 변환 과정에서 데이터 범위를 조절할 수 있으나 성능에 영향을 줄 수 있기 때문에 최적의 성능을 위한 파라미터로 자동화 할 수 있다. 그림 2는 제안하는 프레임워크에서 컴포넌트 적응부에 적용될 수 있는 파라미터 입력 GUI를 보여준다. 적응부에서 입력받을 수 있는 파라미터는 크게 인코더 및 디코더 유형과 이미지 스케일, 데이터 범위이다. 데이터 범위의 경우 그림 3과 같이 -1부터 1, 0부터 1, 0부터 2, 0부터 5 그리고 0부터 10으로 총 5가지의 방법으로 수행해보았으며 -1부터 1일 때 최적의 성능을 보일 수 있는 것을 확인하였다. 따라서 IoT 개발자에게는 -1부터 1과 같은 범위로 자동화 하여 제공한다.



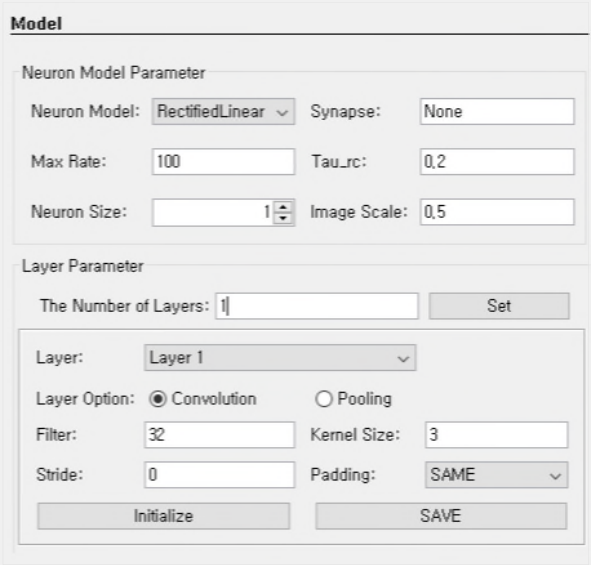
(그림 2) 적응부 파라미터 입력 GUI



(그림 3) 데이터 범위에 따른 수행 결과

3.3 모델부

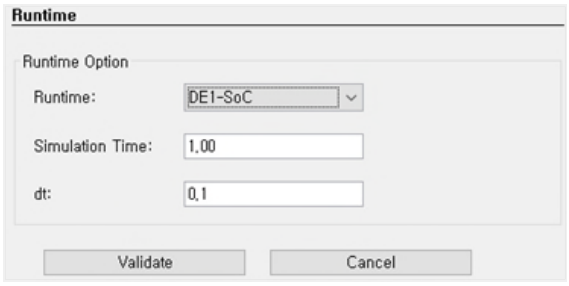
컴포넌트의 모델부는 인공신경망 모델을 위해 성능에 영향을 줄 수 있는 파라미터를 추상화하여 입력받도록 한다. 특히 스파이킹 신경망은 2세대 인공신경망에 비해 복잡한 구조로 되어있어 조절 가능한 파라미터가 많으며 성능에 영향이 없는 파라미터도 있는 것을 확인하였다. 따라서 그림 4와 같이 성능과 직접적 연관으로 조절 가능한 파라미터를 입력받는다. 추상화된 파라미터는 하드웨어에서 동작할 수 있는 뉴런의 사용률인 neuron size, 뉴런이 작동하기 위한 발화율을 보여주는 max rate, 뉴런 출력값의 저주파 통과 필터인 synapse, 막전압 0으로 수렴하는 빠르기를 나타내는 tau rc, 뉴런 출력의 진폭을 나타내는 amplitude이다.



(그림 4) 모델부 파라미터 입력 GUI

3.4 실행부

컴포넌트의 실행부는 IoT 개발자가 선택한 하드웨어를 고려하여 스파이킹 신경망으로 생성된 모델을 최적화 할 수 있다. 먼저, 생성된 모델에 대해 실행 가능한 하드웨어 목록을 보여주고, 수행 가능한 하드웨어가 없다면 시뮬레이션, 외부 서버, 2세대 인공신경망 순서로 추천할 수 있다. 시뮬레이션 수행이 불가능할 경우 외부 서버로 유도를 하게 되는데 이는 하드웨어 및 시뮬레이션으로 수행이 어려울 정도로 규모가 크다고 볼 수 있다. 모델이 외부 서버로 연동되어 수행하기 전에 생성한 모델과 비슷하거나 동일한 외부 모델이 있는지 확인한다. 만약 외부에서 동작 가능한 컴포넌트가 없을 경우에는 2세대 인공신경망으로의 변환을 제안할 수 있으며 최대한 뉴로모픽 아키텍처 기반의 하드웨어 또는 시뮬레이션으로 수행할 수 있도록 중점을 두었다.



(그림 5) 실행부 파라미터 입력 GUI

4. 결 론

본 논문에서는 IoT 개발자를 위해 파라미터 입력만으로 쉽게 자율형 IoT 응용을 개발할 수 있도록 뉴로모픽 하드웨어를 고려한 컴포넌트 모델 최적화 기법을 제안하였다. 뉴로모픽 하드웨어는 스파이킹 신경망을 통해 빠른 속도로 스스로 학습, 인식 및 추론하는 자율형 IoT를 지원할 수 있지만, 생물학적 뉴런에 대한 지식을 요구하기 때문에 IoT 개발자는 어려움이 따른다. 제안 기법을 통해 생성되는 컴포넌트는 인코딩 유형 및 데이터 크기를 자동화할 수 있는 적응부, 모델 성능에 영향받지 않는 파라미터를 추상화하는 모델부, 실행 환경에 최적화된 모델을 제공하는 실행부를 통해 최적화된 모델을 가지도록 생성된다. IoT 개발자는 파라미터 입력만으로 AI 컴포넌트를 생성할 수 있고 플로우 기반의 오픈 IDE에 적용하여 AI와 결합된 자율형 IoT 응용을 쉽게 개발할 수 있다.

Acknowledgement

This work was supported by Institute for information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00708, Integrated Development Environment for Autonomic IoT Applications based on Neuromorphic Architecture).

참 고 문 헌

[1] Yangja Jang. "IoT Platform Technology Review." Communications of the Korean Institute of Information Scientists and Engineers 32.6 (2014): 19-24.

[2] 김대영, 라현정, 김수동. "IoT 디바이스 이질성의 효율적 관리를 위한 프레임워크." 정보과학회논문지 : 소프트웨어 및 응용 41.5 (2014): 353-366.

[3] 전종암, 김내수, 고정길, 박태준, 강호용, 표철식. "IoT 디바이스 제품 및 기술 동향." 한국통신학회지(정보와통신) 31.4 (2014): 44-52.

[4] 박현문, 황태호. "엣지컴퓨팅기술의 변화와 동향." 한국통신학회지(정보와통신) 36.2 (2019): 41-47.

[5] Data Driven Investor,“Movidius NCS (with Raspberry Pi) vs. Google Edge TPU (Coral) vs. Nvidia Jetson Nano — A Quick Comparison.”.[Online]. Available : https://medium.com/datadriveninvestor/movidius-ncs-with-raspberry-pi-vs-818427734d3c (Accessed on Jul. 01 2020).

[6] 박동환. "지능형 협업 IoT 디바이스 플랫폼 동향." 전자파기술 30.3 (2019): 3-9.

[7] 이건명. (2020). 스파이킹 뉴런 모델의 동작과 스파이킹 신경망의 학습. 정보과학회지, 38(2), 8-19.

[8] 정재혁, 김서연, 김재희, 김봉재, 정진만."뉴로모픽 하드웨어 지원 IoT 응용 생성 자동화 도구 개발."한국정보과학회 학술발표논문집.(2020):819-821.